Adversarial Surrogate Losses for Ordinal Regression

Rizal Fathony, Mohammad Bashiri, and Brian D. Ziebart
{rfatho2, mbashi4, bziebart}@uic.edu
Department of Computer Science, University of Illinois at Chicago

UIC COMPUTER SCIENCE

Overview

Ordinal Regression (also known as Ordinal Classification)

- Discrete class labels have an inherent order:
  (e.g., *poor*, *fair*, *good*, *very good*, and *excellent* labels)
- Ordinal loss depends on distance between predicted & actual label
- The absolute error, $|\hat{y} - y|$, is a canonical ordinal regression loss

Existing Models

- Reduce the ordinal regression task to multiple subtasks by:
  - Viewing the problem from the regression perspective
    → learn a regression function and a set of thresholds; or
  - Taking a classification perspective
    → use tools from cost-sensitive classification
- Ordinal regression loss: non-convex and non-continuous
  → Surrogate losses for ordinal regression need to be employed
  → Constructed by transforming binary surrogate losses

Our Approach

1. Robust prediction: *what predictor best minimizes absolute error in the worst case, given partial knowledge of the conditional label distribution?*
2. Surrogate losses that realize this adversarial predictor for:
   (1) thresholded regression representation, or
   (2) multiclass representation
3. Enjoys the guarantee of Fisher consistency
4. Performs competitively with linear kernel, and significantly better than state-of-the-art models with Gaussian kernels

Related Works

Support Vector Machines for Ordinal Regression:

- Extend hinge loss to ordinal regression problems
  A. Threshold Methods (Sashua & Levin, '03; Chu & Keerthi, '05; Rennie & Srebro, '05)
     1. All Threshold (also called SVORIM):
        $\text{loss}_{\text{AT}}(\hat{f}, y) = \sum_{k=1}^{y-1} \delta(-(\theta_k - \hat{f})) + \sum_{k=y}^{|\mathcal{Y}|} \delta(\theta_k - \hat{f})$
     2. Immediate Threshold (also called SVOREX):
        $\text{loss}_{\text{IT}}(\hat{f}, y) = \delta(-(\theta_{y-1} - \hat{f})) + \delta(\theta_y - \hat{f})$
  B. Reduction Framework (Li & Lin, 2007)
     - Create $|\mathcal{Y}| - 1$ weighted extended samples for each training sample
     - Run weighted binary classification on the extended samples
  C. Cost Sensitive Classification Methods (Lin, 2008, 201; Tu & Lin, 2010)
     - CS-OVA, CS-OVO, CS-OSR (one sided regression)

Adversarial Prediction Games (Asif et al. 2015)

- Two player zero-sum games:
  1) Adversarial player: controls conditional label distribution $\check{P}(\check{y}|\mathbf{x})$
     → must approximate training data, but otherwise maximize expected loss
  2) Estimator player: controls $\hat{P}(\hat{y}|\mathbf{x})$ and seeks to minimize expected loss
  - Formulation:

$$\min_{\hat{P}(\hat{y}|\mathbf{x})} \max_{\check{P}(\check{y}|\mathbf{x})} \mathbb{E}_{\mathbf{X}\sim P;\hat{Y}|\mathbf{X}\sim\hat{P};\check{Y}|\mathbf{X}\sim\check{P}} \left[\text{loss}(\hat{Y}, \check{Y})\right] \text{ such that: } \mathbb{E}_{\mathbf{X}\sim P;\check{Y}|\mathbf{X}\sim\check{P}}[\phi(\mathbf{X}, \check{Y})] = \tilde{\phi}.$$

  - Feature moments $\tilde{\phi} = \mathbb{E}_{\mathbf{X},Y\sim\tilde{P}}[\phi(\mathbf{X}, Y)]$, are measured from training data
- For ordinal regression, it reduces to an optimization convex in θ:

$$\underbrace{\min_{\mathbf{w}} \sum_i \overbrace{\max_{\check{\mathbf{p}}_{\mathbf{x}_i}} \min_{\hat{\mathbf{p}}_{\mathbf{x}_i}} \hat{\mathbf{p}}_{\mathbf{x}_i}^T \mathbf{L}'_{\mathbf{x}_i,\mathbf{w}} \check{\mathbf{p}}_{\mathbf{x}_i}}^{\text{zero-sum game}}}_{\text{convex optimization of } \mathbf{w}} \; ; \; \mathbf{L}'_{\mathbf{x}_i,\mathbf{w}} = \begin{bmatrix} f_1 - f_{y_i} & \cdots & f_{|\mathcal{Y}|} - f_{y_i} + |\mathcal{Y}| - 1 \\ f_1 - f_{y_i} + 1 & \cdots & f_{|\mathcal{Y}|} - f_{y_i} + |\mathcal{Y}| - 2 \\ \vdots & \ddots & \vdots \\ f_1 - f_{y_i} + |\mathcal{Y}| - 1 & \cdots & f_{|\mathcal{Y}|} - f_{y_i} \end{bmatrix}$$

  where: $\mathbf{w}$ is the Lagrangian model parameter, and $f_j = \mathbf{w} \cdot \phi(\mathbf{x}_i, j)$
- Inner zero-sum game can be solved using a linear program

Adversarial Surrogate Losses

Theorem *An adversarial ordinal regression predictor is obtained by choosing parameters* $\mathbf{w}$ *that minimize the empirical risk of the surrogate loss function:*

$$AL^{ord}_{\mathbf{w}}(\mathbf{x}_i, y_i) = \max_{j,l\in\{1,\dots,|\mathcal{Y}|\}} \frac{f_j + f_l + j - l}{2} - f_{y_i} = \max_j \frac{f_j + j}{2} + \max_l \frac{f_l - l}{2} - f_{y_i},$$

where $f_j = \mathbf{w} \cdot \phi(\mathbf{x}_i, j)$ *for all* $j \in \{1, \dots, |\mathcal{Y}|\}$

Feature representations:

$$\phi_{th}(\mathbf{x}, y) = \begin{pmatrix} y\mathbf{x} \\ I(y \le 1) \\ I(y \le 2) \\ \vdots \\ I(y \le |\mathcal{Y}| - 1) \end{pmatrix} ; \text{ and } \quad \phi_{mc}(\mathbf{x}, y) = \begin{pmatrix} I(y = 1)\mathbf{x} \\ I(y = 2)\mathbf{x} \\ I(y = 3)\mathbf{x} \\ \vdots \\ I(y = |\mathcal{Y}|)\mathbf{x} \end{pmatrix}$$

Thresholded regression surrogate loss: $\text{AL}^{\text{ord-th}}$

$$\text{AL}^{\text{ord-th}}(\mathbf{x}_i, y_i) = \max_j \frac{j(\mathbf{w} \cdot \mathbf{x}_i + 1) + \sum_{k\ge j} \theta_k}{2} + \max_l \frac{l(\mathbf{w} \cdot \mathbf{x}_i - 1) + \sum_{k \ge l} \theta_k}{2} - y_i \mathbf{w} \cdot \mathbf{x}_i - \sum_{k \ge y_i} \theta_k.$$

- $\text{AL}^{\text{ord-th}}$ is based on averaging the thresholded label predictions for potentials $\mathbf{w} \cdot \mathbf{x}_i + 1$ and $\mathbf{w} \cdot \mathbf{x}_i - 1$

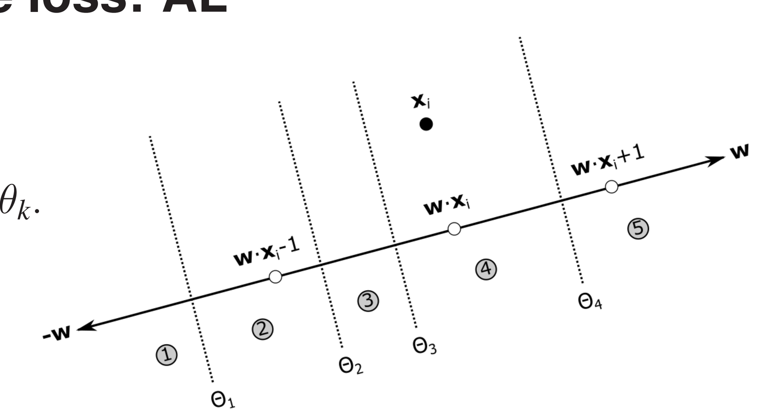

Figure: Surrogate loss calculation for $\text{AL}^{\text{ord-th}}$

Multiclass ordinal surrogate loss: $\text{AL}^{\text{ord-mc}}$

$$\text{AL}^{\text{ord-mc}}(\mathbf{x}_i, y_i) = \max_{j,l\in\{1,\dots,|\mathcal{Y}|\}} \frac{\mathbf{w}_j \cdot \mathbf{x}_i + \mathbf{w}_l \cdot \mathbf{x}_i + j - l}{2} - \mathbf{w}_{y_i} \cdot \mathbf{x}_i$$

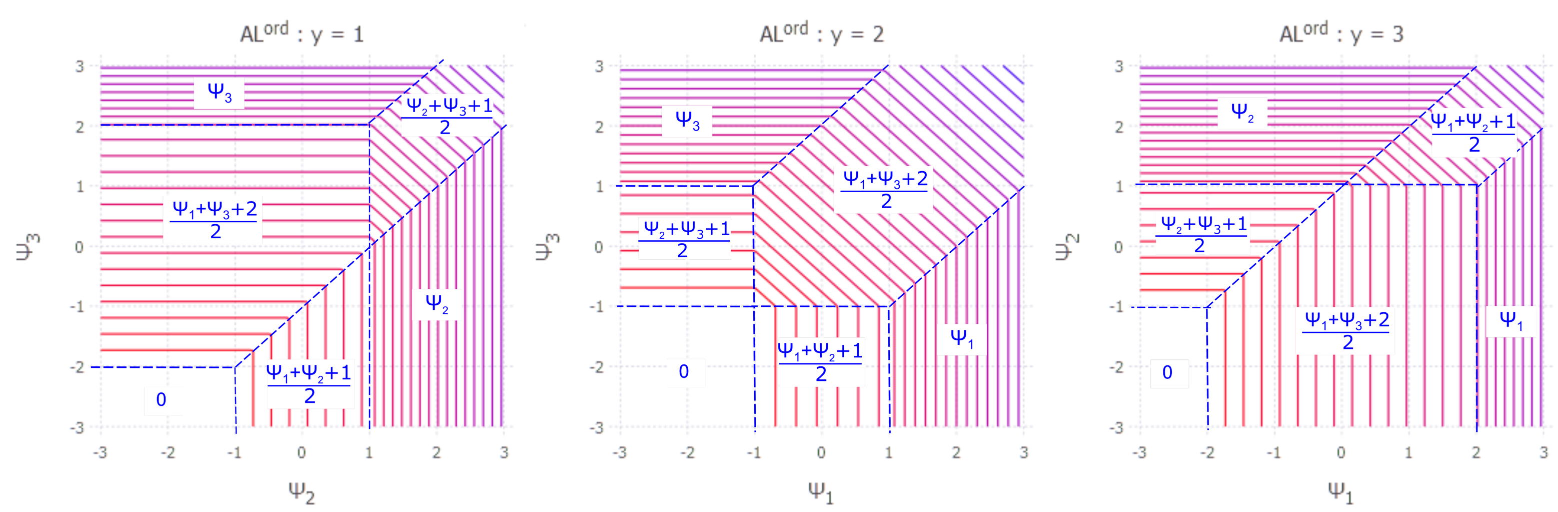

Figure: Loss function contour plots of $\text{AL}^{\text{ord-mc}}$ over the space of potential differences $\psi_j \triangleq f_j - f_{y_i}$ for three classes prediction when the true label is $y_i = 1$, $y_i = 2$, and $y_i = 3$

Fisher Consistency

- Minimizing a Fisher consistent loss yields the Bayes optimal decision boundary given the true distribution, $P(x, y)$
- Ordinal Regression: it requires $\text{argmax}_j f_j^*(\mathbf{x}) \subseteq \text{argmin}_j \sum_y P_y\, |j - y|$, where $P_j \triangleq P(Y = j|\mathbf{x})$ and $\mathbf{f}^*$ is the minimizer of $\mathbb{E}\left[\text{loss}_{\mathbf{f}}(\mathbf{X}, Y)|\mathbf{X} = \mathbf{x}\right]$
- The minimizer of $\mathbb{E}\left[\text{AL}^{\text{ord}}_{\mathbf{f}}(\mathbf{X}, Y)|\mathbf{X} = \mathbf{x}\right]$ satisfies the *loss reflective* property, i.e., it complements the absolute error
- Examples: $[0, -1, -2]^{\text{T}}$, $[-1, 0, -1]^{\text{T}}$ and $[-2, -1, 0]^{\text{T}}$ for three-class problems, and $[-3, -2, -1, 0, -1]$ for five-class problems
- Minimizing over $\mathbf{f}$ that satisfy the loss reflective property is equivalent to finding the Bayes optimal response

Optimization

Primal Optimization using Stochastic Averaged Gradient (SAG)

- SAG (Schmidt et.al, 2013, 2015) uses the gradient of each example from the last iteration it was selected to take a step
- Naïve implementation SAG requires gradient storage
- For AL^{ord}, storage requirement can be drastically reduced by just storing a pair of number, $(j^*, l^*) = \text{argmax}_{j,l\in\{1,\dots,|\mathcal{Y}|\}} \frac{f_j + f_l + j - l}{2}$, rather than the gradient for each sample

Dual Optimization using Quadratic Programming (QP)

- Constrained QP of AL^{ord} plus L2 regularization

$$\min_\theta \frac{1}{2}\|\theta\|^2 + \frac{C}{2}\sum_{i=1}^n \xi_i + \frac{C}{2}\sum_{i=1}^n \delta_i$$

subject to: $\xi_i \ge \theta \cdot \phi(\mathbf{x}_i, j) - \theta \cdot \phi(\mathbf{x}_i, y_i) + j \quad \forall i \in \{1, \dots n\}; j \in \{1, \dots, |\mathcal{Y}|\}$
$\delta_i \ge \theta \cdot \phi(\mathbf{x}_i, j) - \theta \cdot \phi(\mathbf{x}_i, y_i) - j \quad \forall i \in \{1, \dots n\}; j \in \{1, \dots, |\mathcal{Y}|\}$

- Dual QP formulation

$$\max_{\boldsymbol{\alpha},\boldsymbol{\beta}} \sum_{i,j} j(\alpha_{i,j} - \beta_{i,j}) - \frac{1}{2}\sum_{i,j,k,l} (\alpha_{i,j} + \beta_{i,j})(\alpha_{k,l} + \beta_{k,l})(\phi(\mathbf{x}_i, j) - \phi(\mathbf{x}_i, y_i)) \cdot (\phi(\mathbf{x}_k, l) - \phi(\mathbf{x}_l, y_k))$$

subject to: $\alpha_{i,j} \ge 0; \beta_{i,j} \ge 0; \sum_j \alpha_{i,j} = \frac{C}{2}; \sum_j \beta_{i,j} = \frac{C}{2}; i, k \in \{1, \dots, n\}; j, l \in \{1, \dots, |\mathcal{Y}|\}$

- Dual QP only depends on dot products
- Enables efficient rich feature expansion using kernel trick

Experiments and Results

Table: The average of the mean absolute error (MAE) for each model. Bold numbers in each case indicate that the result is the best or not significantly worse than the best.

Dataset	Threshold-based models				Multiclass-based models				
	$\text{AL}^{\text{ord-th}}$	RED^{th}	AT	IT	$\text{AL}^{\text{ord-mc}}$	RED^{mc}	CSOSR	CSOVO	CSOVA
diabetes	**0.696**	**0.715**	0.731	0.827	**0.629**	**0.700**	**0.715**	0.738	0.762
pyrimidines	0.654	0.678	0.615	0.626	**0.509**	0.565	**0.520**	0.576	**0.526**
triazines	**0.607**	0.683	0.649	0.654	0.670	0.673	0.677	0.738	0.732
wisconsin	**1.077**	**1.067**	**1.097**	1.175	1.136	1.141	1.208	1.275	1.338
machinecpu	**0.449**	**0.456**	**0.458**	**0.467**	0.518	0.515	0.646	0.602	0.702
autompg	**0.551**	**0.550**	**0.550**	0.617	0.599	0.602	0.741	0.598	0.731
boston	0.316	**0.304**	**0.306**	**0.298**	0.311	0.311	0.353	**0.294**	0.363
stocks	0.324	0.317	0.315	0.324	0.168	0.175	0.204	**0.147**	0.213
abalone	0.551	0.547	0.546	0.571	**0.521**	**0.520**	0.545	0.558	0.556
bank	0.461	0.460	0.461	0.461	**0.445**	**0.446**	0.732	0.448	0.989
computer	0.640	0.635	0.633	0.683	**0.625**	**0.624**	0.889	0.649	1.055
calhousing	1.190	1.183	1.182	1.225	1.164	**1.144**	1.237	1.202	1.601
average	0.626	0.633	0.629	0.661	0.613	0.618	0.706	0.652	0.797
# bold	5	5	4	2	5	5	2	2	1

- Experiments with Linear Kernel
  - Competitive performance of AL^{ord} compared to baselines on thresholded and multiclass representations
  - AL^{ord} has a slight advantage on the average accuracy
- Experiments with Gaussian Kernel
  - Provides access to much richer feature spaces
  - $\text{AL}^{\text{ord-th}}$ is significantly better than SVORIM (all-threshold model) and SVOREX (immediate-threshold model)

Table: The average of MAE for models with Gaussian kernel.

Dataset	$\text{AL}^{\text{ord-th}}$	SVORIM	SVOREX
diabetes	**0.696**	**0.665**	**0688**
pyrimidines	**0.478**	0.539	0.550
triazines	**0.609**	**0.612**	**0.604**
wisconsin	1.090	1.113	**1.049**
machinecpu	**0.452**	0.652	0.628
autompg	**0.529**	0.589	0.593
boston	**0.278**	0.324	0.316
stocks	**0.103**	**0.099**	**0.100**
average	0.531	0.574	0.566
# bold	7	3	4

Acknowledgments: This research was supported as part of the Future of Life Institute (futureoflife.org) FLI-RFP-AI1 program, grant#2016-158710 and by NSF grant RI-#1526379.